This course has already ended.

« 6. Week: Introduction to the project Course materials 6.2 Getting started with data »

CS-E4730 / 6. Week: Introduction to the project / 6.1 The project

The project

The purpose of the project is for you to try out the skills you learned during the course to perform interesting data analysis for a real computational social science problem. We hope you will find the project the most fun and inspiring part of the course!

During the project, you will work with data from Wikipedia pages of public figures. Wikipedia is a fascinating case study for CSS since it comprises social data on two layers: 1. it contains descriptions of individuals and their significant connections to others, and 2. it results from a collaborative editing process.

Workload, length, and depth of the project

The nominal amount of time we expect students to spend on the project is 30 hours. Depending on the student and the project's ambition level, this time might vary. However, if you spend significantly more than 30 hours, you are most likely trying to do too much.

Structure of the project

We have divided the project into four tasks – Tasks A-D. Every student must complete tasks A and B and *choose* either task C or task D. Tasks A and B give 100 points each, and tasks C or D are worth 200 points. Therefore, the total number of points you can earn for the project is 400. Please do not complete both Tasks C and D since we will only grade the first of the two you report in the project submission. The division of the total points allocated for each task into the sub-tasks is indicated below in the task descriptions.

Finally, the project should be submitted for peer review. Completing the peer review for other students' projects counts for 70 additional points. We will provide further instructions on the peer review process and grading later.

Task A: Review a related paper (100 points)

Review the paper – Women through the glass ceiling: gender asymmetries in Wikipedia
https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-016-0066-4

Please read the paper carefully and write 1-2 paragraphs each to summarize the following in your words:

- Primary goal of the paper (30 points)
- Main methods (20 points)
- Key findings (30 points)
- Brief description of one follow-up analysis you could have done based on the paper if you had unlimited access to data and compute resources (20 points)

Task B: Initial data analysis (100 points)

Download the pre-fetched data here politicians.json and the template notebook here project-wikipedia.ipynb to perform the preliminary data analysis outlined in the notebook. The notebook also contains the division of points for the different sub-tasks. The goal of this task is to perform some initial data analysis that would also set you up for conducting further analysis in Task C if you so choose to do.

Task C: Additional data analysis (200 points)

For this task, please feel free to test out your own analysis ideas or focus on things you find most interesting. You can replicate ideas from the source article or develop your own. Here are some ideas that you can try out or just be inspired by:

- Test if there is homophily related to gender in the politician network. Find the number of links between the different genders and compare it to the baseline case if nodes would have been randomly linking to each other (while keeping the degrees of the nodes fixed). You estimate the number of links between genders in the baseline case either via a simple pen-and-paper computation or by randomising the network explicitly.
- Replicate the analysis of comparing if females/males are more likely to be "local heroes," i.e., if it is true that females are more likely to be in a larger number of language editions than males. Try out different ways of operationalizing this.
- Explore how the language used to describe men and women in their summary texts differs. You could do this by counting gendered words (e.g., she, her, wife) or comparing the usage of terms related to personal life (e.g., father, daughter, married) with professional terms' usage (e.g., leader, president, politician).
- Visualise the network structure between the politicians. You can color nodes with different attributes or take subsamples. You can also attempt clustering the network and inspecting the clusters.
- Find the most central nodes in the network. You could analyze how centrality is related to gender and other attributes.
- Collect more data on other professions, or focus on a different language edition, a specific time era, or nationality. You can repeat your previous experiments on this new data or compute simpler statistics over different subsets of people. Be careful to avoid collecting too much data, as collecting and analyzing data can take a long time.

Note that the data will contain errors and missing data, so you must deal with them in your analysis.

Please remember this is a course project, and you are not producing a research article. So don't hold yourself to the standard of the article you read at the beginning of the project. Research articles can take hundreds or thousands of hours of work from trained scientists, and you have only 30!

- Feel free to cut corners and mention the limitations in your report.
- Report negative results: if you tried something that didn't work, you can still report it.
- Feel free to simplify your analysis, especially compared to the published articles.
- You can choose to do one idea very carefully or prototype several simple ideas.

What you should submit at the end:

- One paragraph explaining what you have done, including the problem statement and brief method description (60 points)
- One paragraph highlighting your main results (140 points)
- Any figures or tables you might need to illustrate your results
- Also, submit your code as a pdf, but please make sure that understanding the results doesn't require reading through your code

Task D: Additional paper reviews (200 points)

Part 1:

For paper in Task A, write the following additional paragraph: - Critical examination of analysis of the dataset for obtaining the results, with respect to the big data characteristics discussed in lecture 3. For instance, you can discuss whether the dataset used by the authors can be considered "always on" or "non reactive", or comment on how representative the data might be, or if any sensitive data is being utilized. You could also briefly discuss if the authors discuss any of the data's limitations in the paper. (50 points)

Part 2:

He Says, She Says: Conflict and Coordination in Wikipedia https://dl.acm.org/doi/pdf/10.1145/1240624.1240698 (alternative link accessible outside of Aalto network http://pensivepuffin.com/dwmcphd/syllabi/info447_wi14/readings/04-ConflictInCollaborations/kittur.HeSaysSheSays.CHI07.pdf)

Please read the paper carefully and write 1-2 paragraphs each to summarize the following in your words:

- Primary goal of the paper (20 points)
- Main methods (20 points)
- Key findings (30 points)
- Brief description of one follow-up analysis you could have done based on the paper if you had unlimited access to data and compute resources (20 points)
- Critical examination of analysis of the dataset for obtaining the results, with respect to the big data characteristics discussed in lecture 3. For instance, you can discuss whether the dataset used by the authors can be considered "always on" or "non reactive", or comment on how representative the data might be, or if any sensitive data is being utilized. You could also briefly discuss if the authors discuss any of the data's limitations in the paper. (50 points)

Project submission instructions

Please return the parts A and C or D as a **single pdf**. More detailed instructions on submitting the project will be given later.

Please **do not copy paste directly** from the papers or online resources. Please write your reviews and project reports in your own words and **cite any online or offline resources** you might have used. And feel free to write your responses in succinct, point-wise form.

« 6. Week: Introduction to the project Course materials 6.2 Getting started with data »